Enhancing Remote Sensing Semantic Segmentation Accuracy and Efficiency through Transformer and Knowledge Distillation

Kang Zheng, Yu Chen, Jingrong Wang, Zhifei Liu, Shuai Bao, Jiao Zhan, Nan Shen

***Abstract*—In semantic segmentation tasks, the transition from convolutional neural networks (CNN) to transformers is driven by the latter's superior ability to capture global semantic information in remote sensing images. However, most transformer methods face challenges such as slow inference speed and limitations in capturing local features. To address these issues, this study designs a hybrid approach that integrates knowledge distillation with a combination of CNN and transformer to enhance semantic segmentation in remote sensing images. First, this paper proposes the Dual-Path Convolutional Transformer Network (DP-CTNet) with a dual-path structure to leverage the strengths of both CNN and transformers. It incorporates Feature Refinement Module to optimize the transformer's feature learning, and Feature Fusion Module to effectively merge CNN and transformer features, preventing the insufficiently learning of local features by the transformer. Then DP-CTNet serves as the teacher model, and pruning and knowledge distillation are employed to create Efficient DP-CTNet (EDP-CTNet) with superior segmentation speed and accuracy. Angle Knowledge Distillation (AKD) is proposed to enhance the feature migration learning of DP-CTNet during knowledge distillation, leading to improved EDP-CTNet performance. Experimental results demonstrate that DP-CTNet thoroughly combines the respective advantages of CNN and Transformer, maintaining local detail features while learning extensive sequential semantic information. EDP-CTNet not only delivers impressive segmentation speed but also exhibits excellent segmentation accuracy following AKD training. In comparison to other models, the two models proposed in this paper notably distinguish themselves in terms of accuracy and result visualization.**

***Index Terms*—convolutional neural network, remote sensing, semantic segmentation, transformer.**

This work was supported by the Natural Science Foundation of Jiangsu Province under Grant BK20220367. (Corresponding author: Yu Chen and Jingrong Wang).

Kang Zheng is with The College of Geography and Environment Science, Henan University, Kaifeng 475004, China (email: zkang0211@whu.edu.cn).

Yu Chen, Nan Shen are with the School of Geomatics and Technology, Nanjing Tech University, Nanjing 211816, China (email: ychen7121@njtech.edu.cn; nanshen@njtech.edu.cn).

Jingrong Wang, Jiao Zhan are with the GNSS Research Center, Wuhan University, Wuhan 430079, China (email: wangjingrong@whu.edu.cn; zhanjiao1994@whu.edu.cn).

Zhifei Liu, Department of Aerospace and Geodesy, Technical University of Munich, 80333 Munich, Germany(e-mail:lzfgiser2001@163.com)

Shuai Bao, China and Research Center of Geospatial Big Data Application, Chinese Academy of Surveying and Mapping, Beijing 100036, China(baogis@163.com)

I. Introduction

The rapid evolution of Earth observation technology has brought about significant improvements in the spatial and temporal resolution of remote sensing imagery [1]. These improved images provide a rich source of data, effectively illustrating the repercussions of human activities on urban landscapes [2]. This wealth of information is highly advantageous for various urban-related applications, including land use classification [3], urban planning [4], and the assessment of urban ecological and environmental impacts [5]. Hence, the effective extraction of land cover information from remote sensing images is a critical challenge that demands attention.

Deep learning holds immense promise for the progress of the remote sensing field [6], with convolutional neural networks (CNNs) making significant contributions to semantic segmentation tasks [7]. Traditional machine learning approaches for semantic segmentation of remote sensing images include support vector machines [8], [9], random forests [10], [11], and conditional random fields [12], [13], among others. However, these methods often depend on handcrafted features and models, whose selection may be limited by expert knowledge and experience, thereby posing challenges in capturing intricate and abstract feature details in high-resolution remote sensing imagery. In contrast, CNNs exhibit superior capabilities in handling the intricate high-level semantic information and diverse terrain features found in high-resolution remote sensing imagery. For instance, U-Net effectively utilizes skip connections to comprehensively capture terrain features [14], and the introduction of the dilated convolutional pyramid pooling module by DeepLab V3Plus has led to a notable boost in network performance [15]. This superior performance can be attributed to CNNs' outstanding feature representation and pattern recognition abilities, thereby solidifying the suitability of CNN-based approaches for semantic segmentation tasks in high-resolution remote sensing imagery [16], [17].

Nonetheless, CNNs require a sequence of pooling and downsampling operations, leading to the loss of a substantial amount of contextual spatial information [18]. At present, CNNs are addressing this issue to some extent by integrating attention mechanisms. These mechanisms establish connections between context and channels, thereby enhancing

the feature learning capacity of network models [19]. However, in high-resolution remote sensing imagery, the characteristics of land cover types become increasingly complex, particularly for highly similar land cover types [20], necessitating more robust model in extracting global context and spatial features. This is essential for further enhancing segmentation accuracy.

Recently, the transformer has emerged as a prominent approach for semantic segmentation tasks. In contrast to CNNs, transformers are entirely built upon self-attention mechanisms, endowing them with more potent capabilities for learning contextual semantic information [21]. Vision Transformer (ViT), which is based on the encoder-decoder structure of transformers, has demonstrated that CNNs are not necessarily required for semantic segmentation tasks, and it has exhibited excellent performance in segmentation tasks [22]. Subsequently, Swin Transformer introduced shifted windows and a hierarchical architecture to enhance model efficiency and flexibility [23], and it has found widespread applications in semantic segmentation of remote sensing images [24], [25]. ST-UNet [26] embeds the Swin Transformer into the classic CNN-based U-Net to capture global contextual semantic information of remote sensing images, aiming to improve the accuracy of land cover segmentation. Zhang *et al.* [27] adopt the Swin Transformer as an encoder to capture long-range information of remote sensing images, designing a CNN-based decoder to restore the size of feature maps. These studies employ CNNs as decoders and design them to be integrated with Swin Transformer as encoders through skip connections, highlighting the robustness of Swin Transformer in remote sensing image semantic segmentation tasks [28]. However, this encoder-decoder combination for CNNs and transformers is not without imperfections. Furthermore, directly coupling the low-order fine-grained detailed feature mapping generated by the encoder with the high-order coarse-grained semantic information generated by the decoder is not appropriate [29]. These approaches can lead to inconsistent feature representation, further resulting in feature loss.

Additionally, the high computational cost of transformers leads to slower inference speeds, which is a current challenge. l Lightweight transformers have gradually emerged as a solution to this problem, such as MobileViT [30], TopFormer [31], and SCAT [32]. In the field of semantic segmentation of remote sensing images, UNetFormer [33], Efficient Transformer [34] and LightFGCNet [35] have all achieved satisfactory results. These methods can be broadly categorized into two approaches: one involves integrating lightweight CNNs as feature extractors directly into the preceding layers, while the other focuses on designing lightweight modules for feature extraction, thereby reducing model complexity and the number of parameters [36]. However, when dealing with high-resolution remote sensing imagery, lightweight transformers often struggle to sufficiently learn complex higher-order semantic information due to their fewer parameters, resulting in segmentation outcomes that are often inferior to non-lightweight methods.

Therefore, to overcome the challenges in integrating CNNs and transformers, as well as to enhance the learning capacity of lightweight transformers, this paper introduces enhancements focusing on precision and efficiency. On one hand, the Dual-Path Convolutional Transformer Network (DP-CTNet) is proposed, aiming to mitigate the inconsistency in feature representation caused by the encoder-decoder structure by integrating CNN and Transformer in a dual-path manner. On the other hand, the DP-CTNet is lightweighted using pruning techniques, resulting in the Efficient DP-CTNet (EDP-CTNet).

Relying exclusively on pruning techniques can adversely affect the model's segmentation performance[37], as it may diminish the learning capacity of EDP-CTNet. Knowledge distillation leverages the transfer of knowledge from a teacher model to improve the performance of lightweight student models [38]. It typically utilizes metric-based learning, where distance-wise distillation loss directs the student model by measuring the discrepancies between the teacher and student features [39]. While such methods capture broad feature differences, they often struggle to convey more detailed knowledge structures, resulting in insufficient knowledge transfer. To address this limitation, this paper introduces an Angle Knowledge Distillation (AKD) loss function, designed to improve knowledge transfer and enable lightweight transformers to effectively learn from high-resolution remote sensing images.

The primary contributions of this paper can be summarized as follows.

1. We construct an innovative model by integrating CNN and transformer with two separate pathways. This architectural design encourages the network to learn global contextual information while mitigating the loss of local details. Furthermore, inspired by BiSeNet [40], this structure supports the lightweight processing of DP-CTNet.
2. In the transformer pathway, we propose the FRM to optimize the output features at various stages. Leveraging multiple pooling techniques, FRM enhances the network's ability to capture global semantic information and guides feature learning. In the process of feature fusion, we design the FFM to effectively combine the features of both CNN and transformer. Given the differing receptive fields of CNN and transformer, our approach involves setting two distinct pathways to obtain corresponding weights, enabling the efficient fusion of semantic information.
3. We propose the AKD and apply it to the pruned DP-CTNet to distill knowledge, resulting in a high-performance EDP-CTNet. Traditional knowledge distillation loss functions calculate using KL divergence [41], but in the case of complex remote sensing information, combining angular attributes may be more effective in conveying relational information to the EDP-CTNet.
4. In comparison to this state-of-the-art approach, both DP-CTNet and EDP-CTNet demonstrate outstanding performance in semantic segmentation of remote sensing images across two datasets. Moreover, EDP-CTNet exhibits superior inference speed compared to transformer-based methods.

The remainder of this paper is organized as follows. Related work to the method presented in this paper is discussed in Section II. The proposed method is introduced in Section III. Relevant details of the experiment can be found in Section IV. Ablation experiments, comparisons with other methods, and efficiency analysis are presented in Section V. Finally, the conclusion is provided in Section VI.

II. Related Work

A. CNN-Based Semantic Segmentation Methods

The introduction of the Fully Convolutional Network marked the pioneering application of CNN structures for end-to-end semantic segmentation problem solving. [42], [43]. Subsequently, U-Net introduced an encoder-decoder architecture tailored for the extraction and reconstruction of distinctive features in semantic entities [14]. Furthermore, the Residual Network tackled the problem of gradient vanishing in deep networks by implementing a sequence of residual blocks, effectively preventing the network from being trapped in local minima [44]. The DeepLab series of networks, through a process of continual refinement, introduced techniques such as atrous spatial pyramid pooling and dilated convolution, significantly enhancing semantic segmentation performance [15], [45]。

In the context of remote sensing image semantic segmentation, CNN-based methods are progressively gaining prominence. Notable contributions include the introduction of the "Semi-Transfer Deep Convolutional Neural Network" by Huang Bo et al., offering innovative solutions to urban land use mapping challenges in remote sensing imagery [46]. Sun Ying et al. have further advanced the field by leveraging CNNs to fuse laser radar data with high-resolution remote sensing imagery for semantic segmentation [47]. However, CNNs exhibit limitations in their capacity to extract contextual information, often leading to feature information loss.

The integration of CNN with attention mechanisms has emerged as an effective approach to addressing this issue. The multiattention network (MANet) [48] tackles context dependency by employing an efficient attention mechanism module. ISANet [49] introduces interlaced sparse self-attention to enhance segmentation efficiency. DFANet [50] proposes a fully-connected attention module for the fusion of CNN features, thereby improving the model's segmentation performance.

However, these methods based on the attention mechanism are obtained by aggregating local features extracted by CNNs to obtain access to global information, rather than modeling the global information directly, which may result in the extracted global information being non-comprehensive.

B. Transformer-Based Semantic Segmentation Methods

The transformer concept first emerged within the realm of natural language processing tasks, where it led to substantial improvements in performance [21], [51]. This architectural approach found extensive application within the domain of computer vision, object detection, semantic segmentation, image enhancement, and more[52], [53], [54]. In the context of semantic segmentation, ViT [55] has already demonstrated reliable performance and displayed remarkable potential. Swin-Unet [56] adopts a purely transformer-based structure along with a U-shaped architecture for the semantic segmentation of medical images. TransUNet [57] employs the transformer as an encoder to capture global contexts and complements this with a CNN-based decoder for the upscaling process. SRCBTFusion-Net [29] adopts an encoder-decoder structure to fuse CNN and transformers, enhancing semantic segmentation performance in remote sensing imagery. DCSwin [24] introduces Swin Transformer as the backbone and designs a densely connected feature aggregation module to restore the original image size. CMTFNet [58] is a CNN and multiscale transformer fusion network designed to extract both local information and global contextual information from remote sensing imagery. While these networks exhibit outstanding segmentation capabilities, they often lack efficiency, requiring significant training and inference times.

To address these efficiency challenges, various lightweight transformer models have been proposed. SegFormer [59], for instance, stands out as a simple yet powerful semantic segmentation model that introduces lightweight multilayer perception for the aggregation of information from different layers. BANet [60] takes advantage of dependency and texture paths to effectively capture long sequential relationships and detailed information, facilitating the rapid extraction of semantic features. MobileViT [30], on the other hand, represents a lightweight and versatile transformer suitable for mobile devices. In a similar vein, Lawin Transformer [61] introduces large window attention spatial pyramid pooling and demonstrates superior performance compared to MaskFormer [62].

Although these lightweight transformer models improve the inference speed to a certain extent, these methods result in a semantic segmentation performance that is inferior to other non-lightweight transformer models because of the reduction in the number of parameters and the depth of the network. images [56]. TransUNet employs the transformer as an encoder to capture global contexts and complements this with a CNN-based decoder for the upscaling process [57]. While these networks exhibit outstanding segmentation capabilities, they often lack efficiency, requiring significant training and inference times.

III. Method

The research framework of this paper, as illustrated in Fig. 1, comprises two main components. The first part involves the construction of DP-CTNet, while the second part focuses on the development of an efficient EDP-CTNet through pruning and knowledge distillation. Detailed descriptions of these components are provided in the subsequent sections.

A. DP-CTNet

As shown in Fig. 1 and Fig. 2, DP-CTNet consists of two pathways: the CNN path and the transformer path. Compared to the U-shaped architecture, this design reduces computational

complexity and enhances the model's execution speed [40]. The CNN path is primarily responsible for extracting local detailed information and is constructed using residual blocks. The transformer path focuses on capturing global contextual information and is built using Swin Transformer blocks. The FRM is employed to optimize the features from the transformer, while the FFM facilitates the comprehensive integration of both pathways.

The Residual Blocks in the CNN path employ skip connections to reduce information loss during the forward propagation process. Additionally, they combine the results of 16x downsampling with 8x downsampling to enhance the model's ability to learn local features of remote sensing objects. In the Transformer path, the Swin Transformer Block consists of window multi-head self-attention (W-MSA) and shifted-window multi-head self-attention (SW-MSA). W-MSA computes self-attention within a fixed window size, while SW-MSA extends this by introducing window shifts. The Transformer path leverages the FRM to aggregate the results of 8x, 16x, and 32x downsampling to improve the model's capacity to capture contextual semantics.

B. Feature Refinement Module

Transformers usually flatten and project image patches into a hierarchical network [63]. However, due to the presence of dense and small-scale objects in remote sensing images, such methods can lead to the loss of fine-grained details and structural information [26]. Each channel serves as a feature detector, focusing on the "meaningful content" in the image [64], [65]. Hence, the FRM designed in this paper integrates

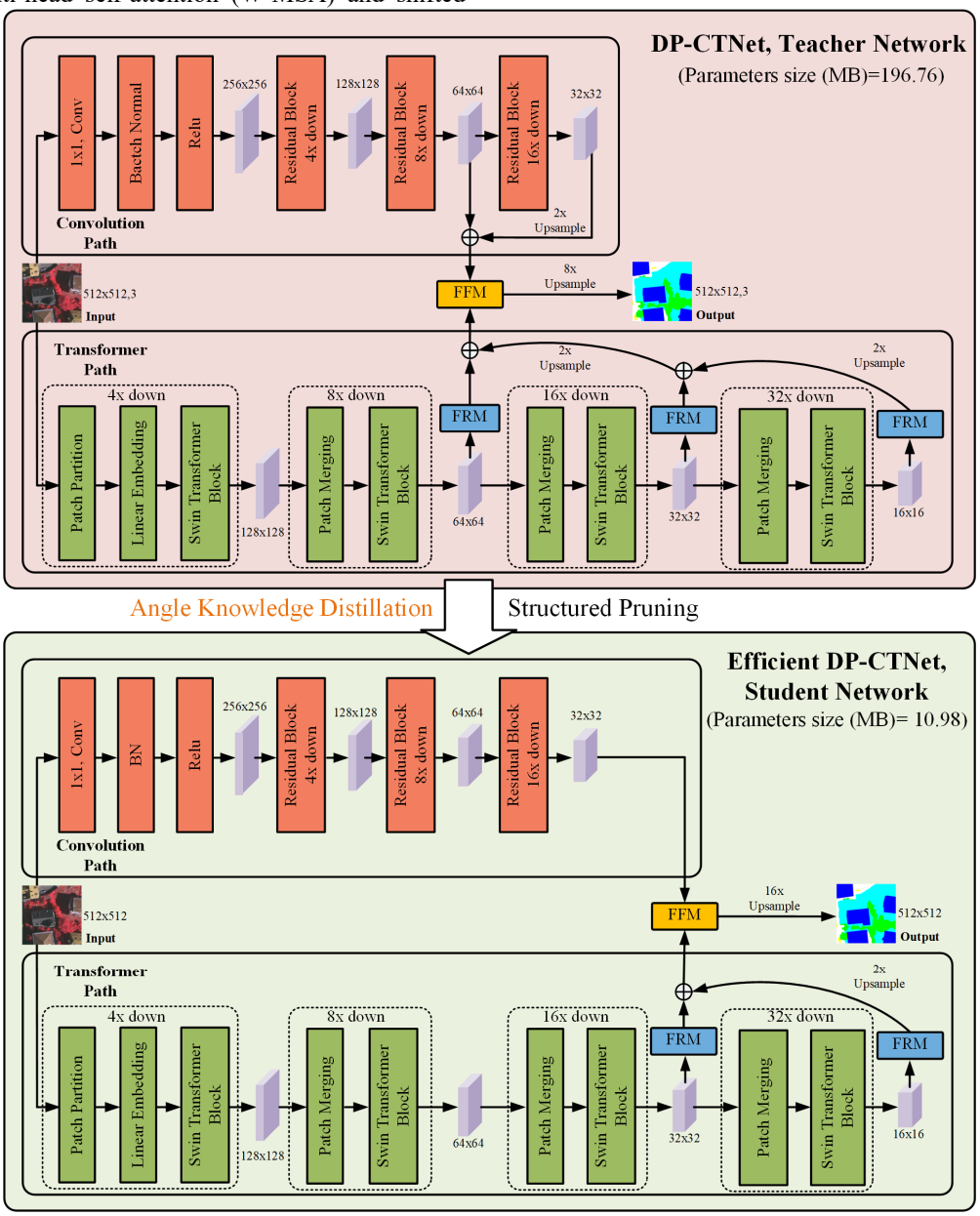


Fig. 1. Research Framework. DP-CTNet model consists of Residual blocks and Swin Transformer blocks, yielding output through 8-fold upsampling with a parameter volume of 196.76MB. In the design of Efficient DP-CTNet, aiming for lightweight and

improved model inference speed, depthwise separable convolutions were employed as replacements, along with a series of structured pruning measures. For instance, some network layers were removed, the channel quantity of the output feature maps from the final two layers in both paths was reduced, and the output resolution was changed to 16-fold upsampling. Ultimately, the model parameter volume was reduced to 10.98MB. Additionally, to prevent potential degradation in segmentation performance in Efficient DP-CTNet, the Angle knowledge distillation method was introduced.

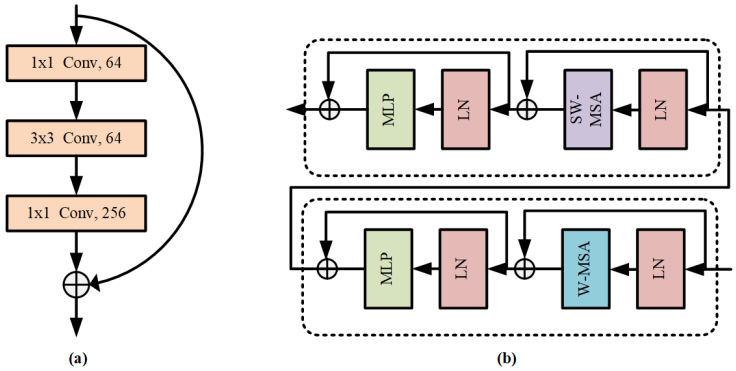


Fig. 2. An Overview of the Residual Block and Swin Transformer Block. (a) Residual Block. (b) Swin Transformer Block.

three pooling strategies to capture channel dependencies. It is incorporated into the downsampling process of the Swin Transformer to optimize the features of each downsampling result, thus addressing the aforementioned issues and enhancing the segmentation of small-scale objects.

The structure of the FRM, as depicted in Fig.3., draws inspiration from the attention mechanism refinement module introduced in BiSeNet[66]. It employs both average pooling and max-pooling to extract features and calculate attention vectors, guiding feature learning. Additionally, it incorporates stripe pooling to facilitate information fusion and explore complex scenes [67]. Specifically, average pooling and max-pooling are applied to the input feature maps to compute global statistical features along the channels. The pooled feature maps are then subjected to convolutional operations to further extract features. A sigmoid function is utilized to map the summed feature maps, generating weights that indicate their importance. Subsequently, the weighted feature maps are fused with the original input feature maps through multiplication. Additionally, the original input feature maps undergo horizontal and vertical strip pooling to aggregate local contextual information. A 1D convolution with a kernel size of 3 is applied to expand and add the pooled feature maps while maintaining consistent dimensions. The resulting feature maps are optimized through 2D convolutional operations. Finally, the results of the strip pooling and the weighted feature maps are combined through addition to yield the feature-optimized output of the Swin Transformer blocks. This hybrid pooling strategy can better achieve a more comprehensive understanding of channel dependencies within the Transformer path.

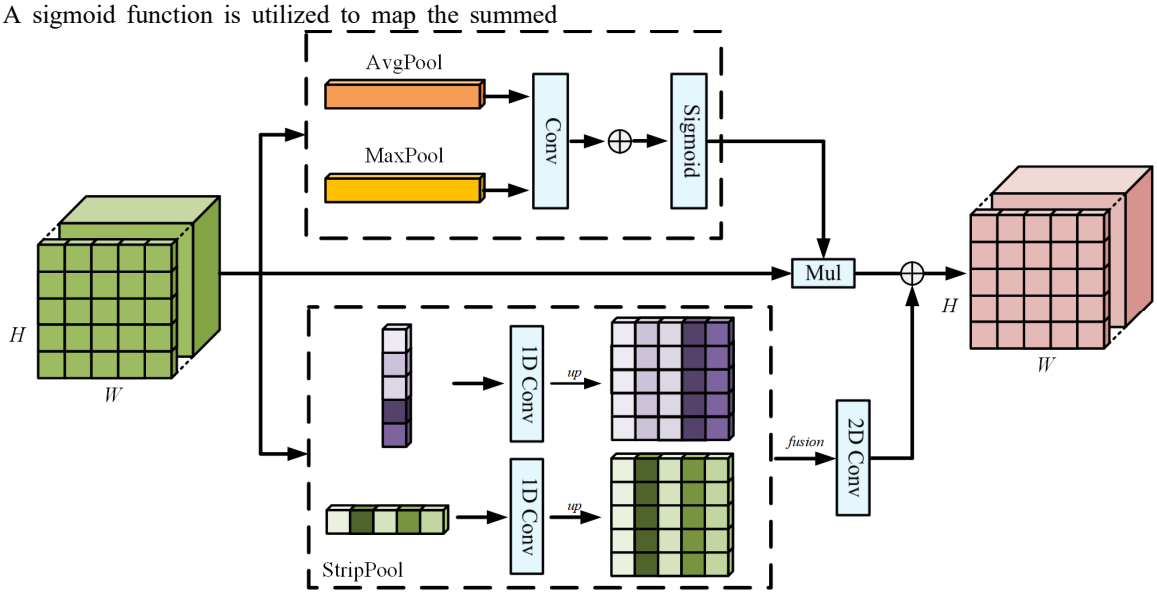


Fig. 3. Structure of the FRM. It optimizes the feature maps of different stages in the transform path by integrating three distinct pooling methods.

C. Feature Fusion Module

There are significant differences between Transformers and CNNs. While CNNs construct receptive fields locally and gradually expand them iteratively, the global interaction mechanism of Transformers allows rapid expansion of receptive fields. Consequently, directly adding the outputs of Transformer and CNN with larger receptive fields does not fully integrate their output feature maps [68], [69]. To mitigate this issue, the FFM is designed with two pathways, as illustrated in Fig. 4.

It separates Transformer and CNN with large receptive fields and combines them with different weights. Specifically, FFM includes a short skip connection scenario and a long skip connection scenario. In the short skip connection scenario, the feature maps from the convolution path and the transformer path are added together, and the resulting feature map is optimized through average pooling and max-pooling. Subsequently, different weights are computed based on the sigmoid function. In the long skip connection scenario, the initial features of CNN and Transformer are multiplied by their respective weights and then added together. The specific calculation formula is as follows.

$$\mathrm{Z} = \mathrm{M}(\mathrm{X}+\mathrm{Y}) \times \mathrm{X} + \big(1 - \mathrm{M}(\mathrm{X}+\mathrm{Y})\big) \times \mathrm{Y}. \qquad (1)$$

Here, X represents the output features of CNN, Y represents the output features of transformer, M represents the channel attention module, and $\mathrm{Z} \in \mathbb{R}^{C\times H\times W}$ represents the fused feature results. It is worth noting that M serves to optimize the features of the $(\mathrm{X}+\mathrm{Y})$ result and guide the network to conduct a soft selection between X and Y.

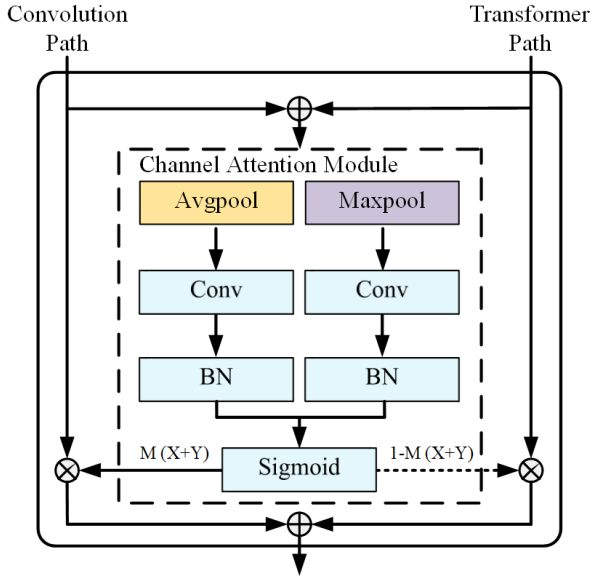


Fig. 4. Structure of the FFM.

In the FFM, it is evident that the combination of long-range features from Transformer and local features from CNN, followed by input into the channel attention module, dynamically selects appropriate weights. This adaptive adjustment facilitates adaptability to feature maps with different receptive fields. Compared to direct feature fusion, this adaptive adjustment method is more advantageous in harnessing the respective strengths of Transformer and CNN.

D. Angle Knowledge Distillation

Large parameter size and slow inference speed are significant drawbacks of transformer, and lightweight transformer with small parameter sizes often struggle to meet segmentation accuracy requirements. Leveraging lightweight techniques such as pruning and knowledge distillation can effectively enhance the performance of lightweight transformer [70]. Traditional knowledge distillation loss functions are divided into two parts: one part involves the Kullback-Leibler divergence [71] between the probability distributions of the teacher model (large parameter transformer) and the student model (lightweight transformer), while the other part concerns the cross-entropy between the output of the student model and the labels [72], [73], [74]. However, when faced with large parameter transformer models and high-resolution remote sensing images, relying solely on traditional knowledge distillation loss functions for transfer is insufficient. Therefore, this paper proposes the incorporation of angle distillation loss on top of the traditional knowledge distillation loss function to enhance the performance of lightweight transformer models by fully utilizing teacher model information.

Specifically, the formula for the traditional knowledge distillation loss function is as follows:

$$L = \alpha L_{soft} + (1-\alpha) L_{hard}. \qquad (2)$$

Here, L represents the knowledge distillation loss function, α is the balancing factor, L_{soft} stands for the Kullback–Leibler divergence between the probability distributions distilled from the teacher model and the student model, and L_{hard} denotes the cross–entropy loss between the output of the student model and the labels. The specific formulas for L_{soft} and L_{hard} are as follows:

$$L_{soft} = -\sum_{j}^{N} p_j^T \log\big(q_j^T\big), L_{hard} = -\sum_{j}^{N} c_j \log\big(q_j^1\big). \quad (3)$$

Here, p_j^T and q_j^T represent the probability distributions distilled from the teacher and student models after applying temperature T, c_j represents the true labels, q_j^1 represents the case with T=1.

The angle distillation loss aims to transmit knowledge by utilizing the angular relationship between the output results of the teacher model and the student model. The specific formula is as follows:

$$L_A = \sum_{(i,j,k)} l_\delta\Big(\psi_A\big(e_t^{ij}, e_t^{kj}\big), \psi_A\big(e_s^{ij}, e_s^{kj}\big)\Big). \qquad (4)$$

In this formula, L_A is the angle distillation loss, $i,\ j, k$ represent the three dimensions of the feature map, t_i, t_k, t_k represent the three dimensional feature map computed by the teacher network, l_δ is the smooth L1 loss function, e_t and e_s are feature vectors of the teacher model and student model, ψ_A is the cosine value of two vectors. For the teacher model, the ψ_A formula is as follows:

$$\psi_A = \cos\big(e_t^{ij}, e_t^{kj}\big). \qquad (5)$$

$$e_t^{ij} = \frac{t_i - t_j}{\|t_i - t_j\|_2}, e_k^{ij} = \frac{t_k - t_j}{\|t_k - t_j\|_2}. \qquad (6)$$

Based on the above formulas, the Angle Knowledge Distillation (AKD) process integrates the traditional knowledge distillation function with the angle distillation loss, following the formula below:

$$L_{AKD} = \alpha L_{soft} + \beta L_{hard} + (1 - \alpha - \beta) L_A. \quad (7)$$

Here, L_{AKD} is the AKD loss function designed in this paper, α and β are adjustable the balancing factors. In the experiments of this paper α is set to 0.6 and β to 0.2.

IV. Experiments Details

A. Datasets

1) Vaihingen Dataset: This dataset comprises 33 orthophotos of varying sizes with a ground sampling distance of 9 cm, covering an area of 1.38 km² in Vaihingen. The orthophotos are 8-bit TIFF files with three bands, including near-infrared, red, and green. Corresponding labels are provided for semantic segmentation. The land cover classes represented in the labels are illustrated in Fig. 5 and consist of six categories: Impervious Surface, Building, Low Vegetation, Tree, Car, and Clutter/Background. The dataset was randomly split into training and testing sets in a 7:3 ratio, with each image cropped to a size of 512×512 pixels.

2) Potsdam Dataset: This dataset comprises 38 high-resolution orthophotos, each with a size of 6000×6000 pixels and a ground sampling distance of 5 cm. It covers an area of 3.42 km² in Potsdam, characterized by complex architectural structures and densely populated areas. Each image contains four bands: near-infrared, red, green, and blue, along with corresponding labels. The land cover categories align with those of the Vaihingen dataset, as shown in Fig. 5. Similar to the Vaihingen dataset, the data was divided into training and testing sets in a 7:3 ratio, and each image was cropped to a size of 512×512 pixels.

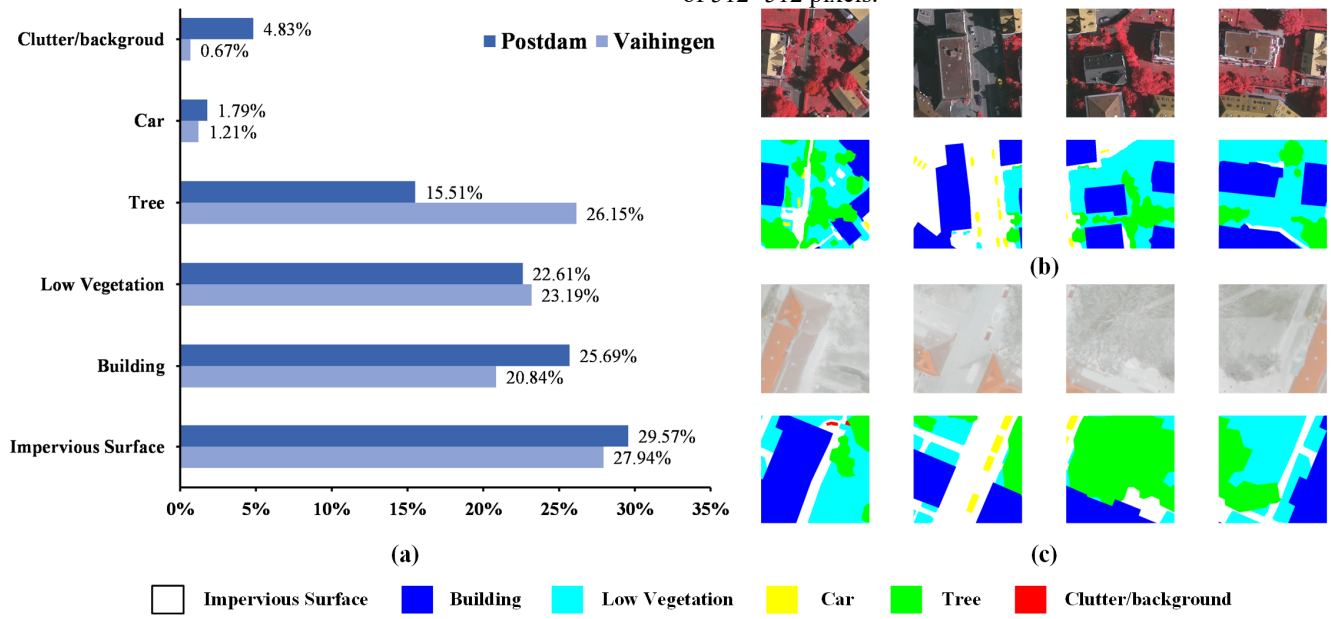


Fig. 5. (a) Proportion of each labels in the Vaihingen and Potsdam datasets. (b) Vaihingen Dataset. (c) Potsdam Dataset.

B. Training Setting

The experiments were conducted within the PyTorch framework. We used the Adam optimizer with a learning rate of 3e-4 and implemented a gradient decay learning strategy. The batch size was set to 8, and the training was carried out for 70 epochs. The hardware used for training was an NVIDIA GeForce RTX 2080 Ti with 11 GB of memory. The training loss function employed was the cross-entropy loss, and the knowledge transfer loss function was L_{AKD}.

C. Evaluation Index

Quantitative evaluation metrics for the segmentation results included the F1-score, overall accuracy (OA), and mean intersection over union (MIoU). These metrics were computed based on the confusion matrix elements, which include true positive (TP), true negative (TN), false negative (FN), and false positive (FP).

The F1-score is a precision evaluation metric based on recall and precision, with the specific formula as follows:

$$F1score = 2 \times \frac{Recall \times Precision}{Recall + Precision}. \quad (8)$$

Where $Recall = \frac{TP}{TP+FN}$ and $Precision = \frac{TP}{TP+FP}$.

OA represents the overall proportion of correctly classified results among all segmentation outcomes, calculated as follows:

$$OA = \frac{TP + TN}{TP + FN + FP + TN}. \quad (9)$$

MIoU measures the intersection over union between ground truth and predicted results, offering an overall assessment of semantic segmentation performance. The formula is as follows:

$$MIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{TP}{FN + FP + TP}. \quad (10)$$

Where N is the number of categories, $i = 1$ denotes the first land cover category.

V. Results and Discussion

A. Ablation Studies

To validate the effectiveness of the proposed FRM and FFM, we conducted ablation experiments on the DP-CTNet using the Vaihingen dataset and Potsdam dataset. We demonstrated improvements in the model through accuracy comparisons and segmentation results.

In Table I, showcasing results from the Vaihingen dataset, when combined with FRM only, the F1-score for Impervious Surface, Building, and Car improved by 0.35%, 0.07%, and 1.99%, respectively, with a noticeable improvement in Car segmentation accuracy. OA and MIoU increased by 0.13% and 0.64%, respectively. FRM significantly contributed to the overall segmentation improvement. Integration with FFM, there were notable enhancements in the segmentation accuracy for all land cover categories and overall segmentation accuracy. Both FFM and FRM played a positive role in remote sensing image semantic segmentation by DP-CTNet. When FFM, and FRM were combined, accuracy improvements were observed compared to the initial model. In comparison to other methods, F1-scores for Impervious Surface, Building, Low Vegetation, and Car exhibited the best performance, with OA and MIoU reaching 86.96% and 73.03%, respectively.

Table I
Comparison of Ablation Experiment Results

Datasets	Models	FRM	FFM	F1-score					OA	MIoU
				Impervious Surface	Building	Low Vegetation	Tree	Car		
Vaihingen	DP-CTNet-FRM-FFM			89.21%	92.01%	84.60%	80.47%	69.42%	86.82%	71.91%
	DP-CTNet-FFM	✓		89.56%	92.08%	84.39%	80.42%	71.41%	86.95%	72.55%
	DP-CTNet-FRM		✓	89.24%	92.11%	84.76%	80.43%	71.88%	86.89%	72.58%
	DP-CTNet	✓	✓	**89.88%**	**92.43%**	84.65%	78.88%	**74.19%**	**86.96%**	**73.03%**
Potsdam	DP-CTNet-FRM-FFM			85.43%	89.46%	74.69%	73.95%	84.12%	82.12%	69.27%
	DP-CTNet-FFM	✓		85.21%	89.25%	76.27%	76.37%	86.42%	83.55%	70.87%
	DP-CTNet-FRM		✓	86.27%	91.78%	75.30%	76.06%	85.91%	84.65%	71.54%
	DP-CTNet	✓	✓	**88.59%**	**93.09%**	**81.02%**	**81.40%**	**87.93%**	**86.32%**	**76.36%**

In Table I, showcasing results from the Potsdam dataset, DP-CTNet demonstrates superior performance across all land cover categories in terms of F1-score. Combined with FRM effectively enhances F1-scores for Low Vegetation, Tree, and Car. FFM integration notably enhances land cover accuracy across various categories. When comparing OA and MIoU, DP-CTNet consistently outperforms combined with FRM and FFM individually. Moreover, the substantial improvement in accuracy observed with different module combinations in the Potsdam dataset, compared to the Vaihingen dataset, is attributed to the significantly larger data volume in Potsdam, providing the model with more training samples and enabling better generalization capabilities.

Based on the results analysed in Fig. 6 for the Vaihingen dataset, the combination of FRM has improved the fragmentation of segmentation results, and it has also enhanced the learning capabilities, particularly for Tree. When combined with FFM, the confusion between Building and Impervious Surface has been notably improved, especially in the red-boxed classification results. The problem of fragmented segmentation has also been alleviated, and in comparison to FRM, FFM shows superior segmentation improvement. DP-CTNet exhibits the best segmentation performance, excelling in both overall continuity and object confusion. This is attributed to the well-designed holistic network architecture that facilitates effective integration between the transformer and CNN.

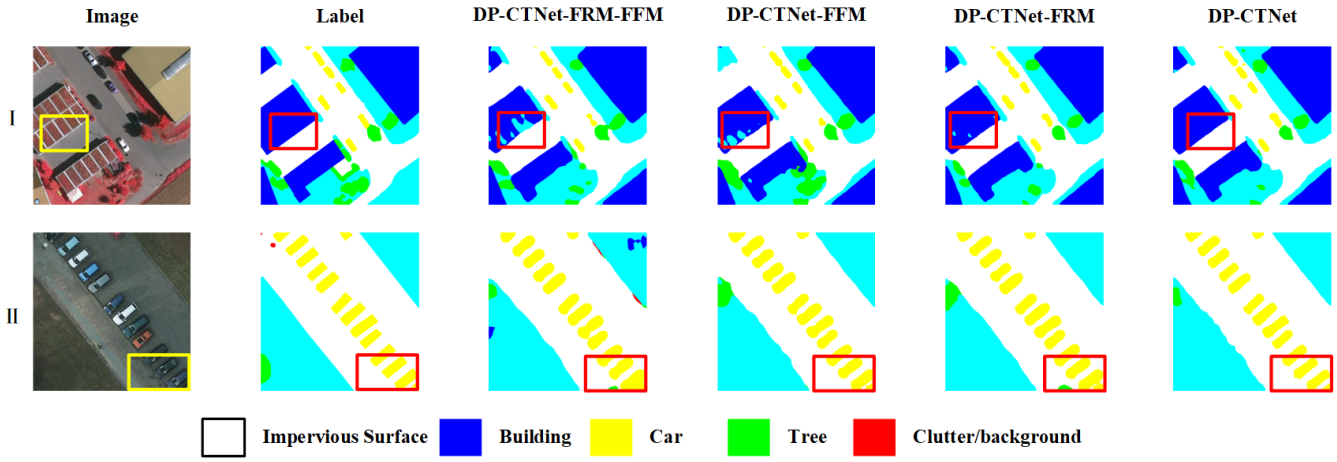


Fig. 6. Segmentation examples of ablation experiments. The rectangular box is the area of highlight comparison. I is the Vaihingen dataset and II is the Potsdam dataset.

Based on the results analysed in Fig. 6 for the Potsdam dataset, it is evident that the incorporation of FRM leads to a noticeable reduction in confusion among land cover classes in segmentation results, with no occurrences of misclassification between Tree and Impervious Surface within the rectangular areas. When combined with FFM, there is an improvement in the edge details of land cover, indicating FFM's effective integration of local features from CNN. However, misclassification issues still persist. In comparison, DP-CTNet shows a certain degree of mitigation in land cover misclassification issues while preserving detailed local information of land cover features.

B. Knowledge Distillation

To demonstrate the effectiveness of AKD, we designed two sets of comparative experiments in the Vaihingen dataset. The first set compared EDP-CTNet trained directly with cross-entropy loss to the same network trained with AKD. The second set involved a comparison between conventional KD and AKD. The comparison was based on different accuracy metrics, with F1-score used for assessing different land covers. All other training parameters were kept consistent.

Analysing the results in Fig. 7, it is evident that the AKD method shows improvements in all accuracy metrics when compared to cross-entropy loss. Although the improvement in F1-score for Tree is not pronounced, there is a substantial 8.39% increase in Car's F1-score. This suggests that AKD facilitates EDP-CTNet in effectively learning the excellent segmentation ability of DP-CTNet for small-sized land covers. Furthermore, the segmentation accuracy of OA and MIoU has also significantly increased.

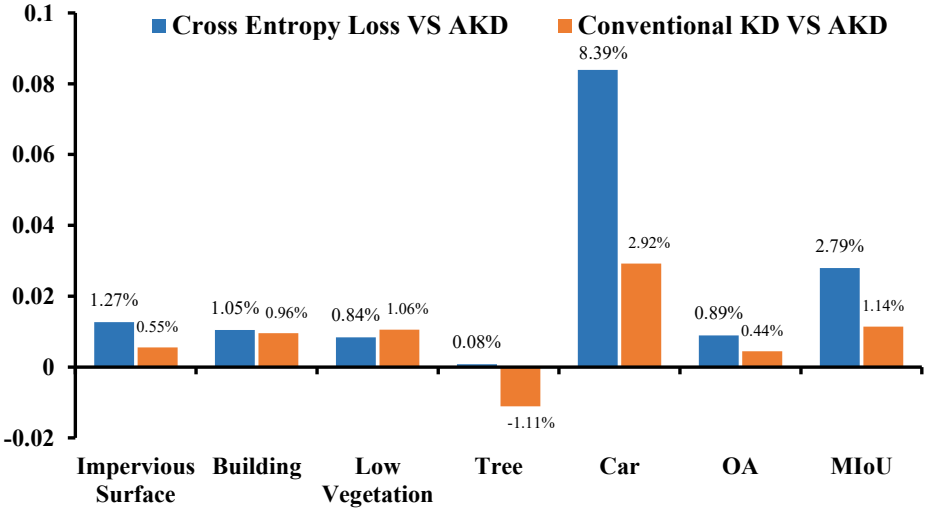


Fig. 7. Comparative Results between AKD and Other Methods.

In comparison to conventional KD, the AKD method doesn't show as significant an improvement in Tree's accuracy. However, AKD outperforms conventional KD in enhancing the segmentation performance of other land covers. This discrepancy might be attributed to the influence of angular properties during the knowledge distillation process, which could affect the learning of Tree. This suggests that considering the height attributes of Tree might be more relevant. In terms of OA and MIoU, the AKD method outperforms conventional KD.

The results indicate that AKD, by incorporating a penalty for angular differences, facilitates the transfer of relational information between the embeddings of training samples. This approach likely enhances the transmission of relational information, thereby enabling EDP-CTNet to achieve superior segmentation performance during training.

C. Comparison between DP-CTNet and other methods

This paper compares DP-CTNet with existing models, including DeepLab V3Plus[45], ISANet[49], CGGLNet[75], LSRFormer[76], MANet[48], TransUNet[57], CMTFNet[58], SRCBTFusion[29], DCSwin[24] and Swin-Unet[56]. DeepLab V3Plus, ISANet, and MANet are CNN-based models. TransUNet and Swin-Unet are transformer-based models. CGGLNet, LSRFormer, CMTFNet, SRCBTFusion, and DCSwin are state-of-the-art methods that fuse Transformer and CNN for semantic segmentation in remote sensing imagery.

1) Result on Vaihingen Dataset: Table II presents the segmentation performance of different models on the Vaihingen dataset. From the table, it can be observed that DP-CTNet has a relatively large parameter size. Furthermore, it outperforms other models in terms of F1-scores for Impervious Surface, Building, and Low Vegetation. However, the performance on Tree segmentation is comparable to other models. Notably, compared to other transformer-based models, DP-CTNet and CGGLNet exhibits a significant advantage in Car segmentation, with a maximum difference of 26.4%, despite being only slightly superior to CMTFNet and SRCBTFusion-Net. This suggests that DP-CTNet effectively combines the strengths of CNN and addresses the transformer's limitations in learning small-scale object information. Regarding OA, DP-CTNet achieves a value of 86.96%, with a less pronounced improvement compared to CGGLNet, DeepLab V3Plus and MANet. In terms of MIoU, DP-CTNet significantly outperforms other models, reaching 73.03%. Overall, DP-CTNet demonstrates superior accuracy, possibly owing to its comprehensive network architecture, which allows it to capture both global contextual information and fine-grained semantic details.

TABLE II
COMPARISON FOR DP-CTNET AND OTHER MODELS ON THE VAIHINGEN DATASET

Models	Parameters Size (MB)	F1-score					OA	MIoU
		Impervious Surface	Building	Low Vegetation	Tree	Car		
DeepLab V3Plus	85.60	89.05%	91.71%	84.29%	80.19%	53.62%	86.48%	68.27%
ISANet	116.83	88.74%	91.41%	83.54%	79.47%	56.53%	85.97%	68.20%
MANet	136.80	89.22%	91.05%	84.55%	80.75%	65.21%	86.54%	70.69%
TransUNet	254.88	85.16%	84.10%	80.08%	78.39%	50.45%	81.97%	62.33%
Swin-Unet	103.67	87.00%	86.54%	82.71%	79.24%	51.97%	83.97%	64.91%
DCSwin	174.07	83.28%	83.14%	78.84%	77.60%	56.23%	80.72%	62.01%
CMTFNet	114.71	88.81%	90.80%	80.68%	69.17%	72.09%	83.24%	67.97%
SRCBTFusion-Net	145.41	87.89%	87.41%	84.41%	81.46%	70.45%	85.47%	70.43%
CGGLNet	62.22	89.70%	90.10%	83.39%	79.93%	76.85%	86.03%	72.76%
LSRFormer	68.03	88.44%	90.11%	83.29%	79.76%	63.00%	85.54%	68.99%
DP-CTNet	196.76	**89.88%**	**92.43%**	**84.65%**	78.88%	74.19%	**86.96%**	**73.03%**

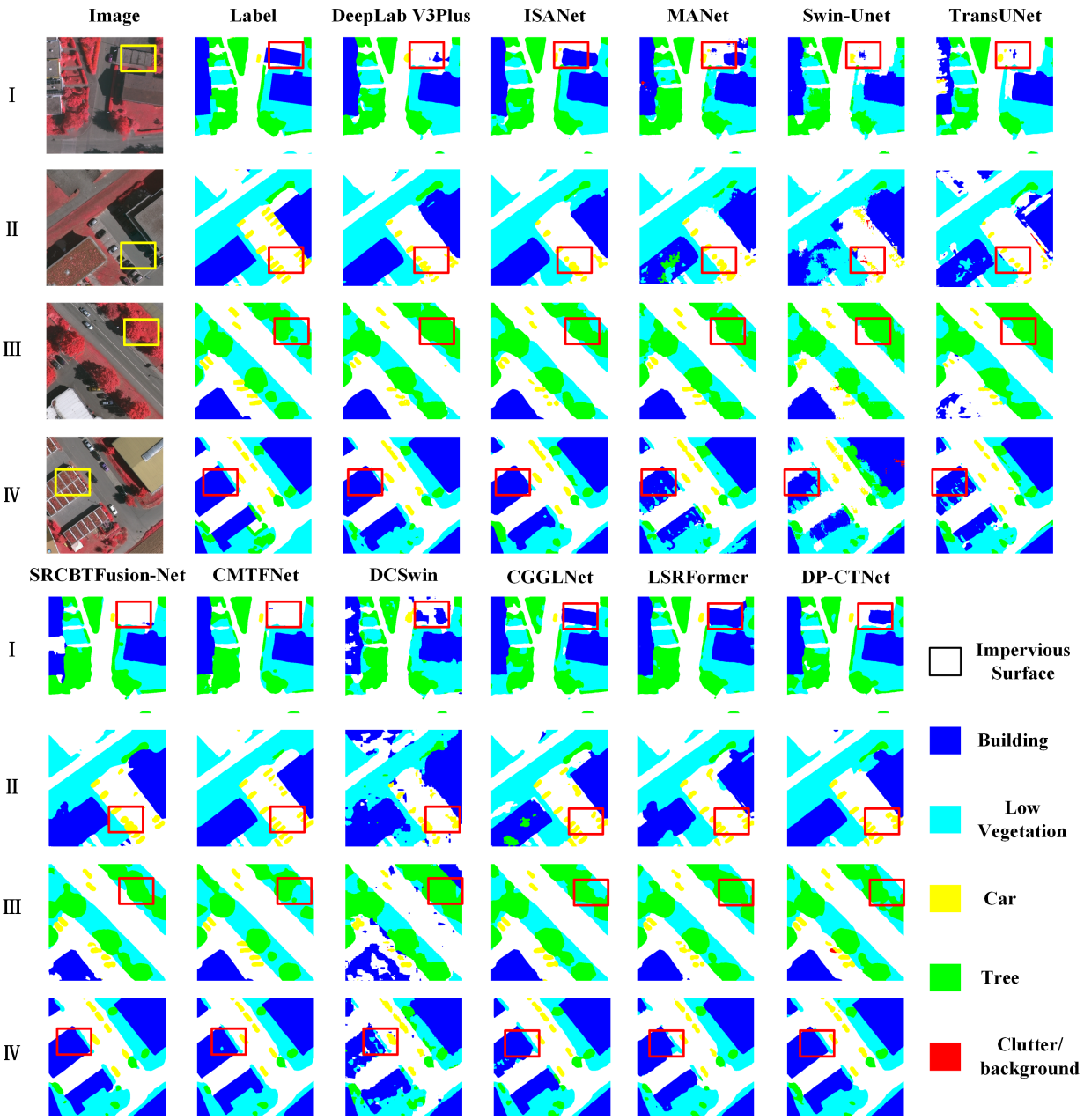


Fig. 3. Examples of Semantic Segmentation Results on the Vaihingen Dataset for DP-CTNet and Other Models. The red and yellow boxes highlight noteworthy areas.

Fig. 3 displays the visual results of different models on the Vaihingen dataset after segmentation. In the first set of images enclosed within the red boxes, it can be observed that DP-CTNet exhibits generally continuous segmentation for Buildings, although not as good as LSRFormer and CGGLNet. Other models display varying degrees of segmentation fragmentation issues. In the second set of images, DP-CTNet's segmentation performance for Cars is notably superior, even in cases with shadow occlusion. DP-CTNet stands out by accurately outlining the right-side Car, only marginally trailing behind CMTFNet's segmentation results. CGGLNet and LSRFormer also have good segmentation results, but mis-segmentation problems occur in other regions. The third set of images, within the red boxes, primarily highlight the segmentation capabilities of different models for Low Vegetation and Trees. DP-CTNet, CMTFNet, SRCBTFusion-Net, ISANet, and MANet demonstrate some level of differentiation between the two, with CMTFNet and DP-CTNet excelling in this aspect. In the fourth set of images, various models exhibit segmentation fragmentation issues for Buildings. SRCBTFusion-Net, DeepLab V3Plus, ISANet, LSRFormer and DP-CTNet demonstrate relatively good overall continuity. However, DeepLab V3Plus, ISANet, and LSRFormer fall short in segmenting Low Vegetation adjacent to Buildings compared to DP-CTNet and SRCBTFusion-Net. Overall, whether in terms of overall continuity or fine-grained segmentation of small-scale objects, DP-CTNet's segmentation results are highly impressive.

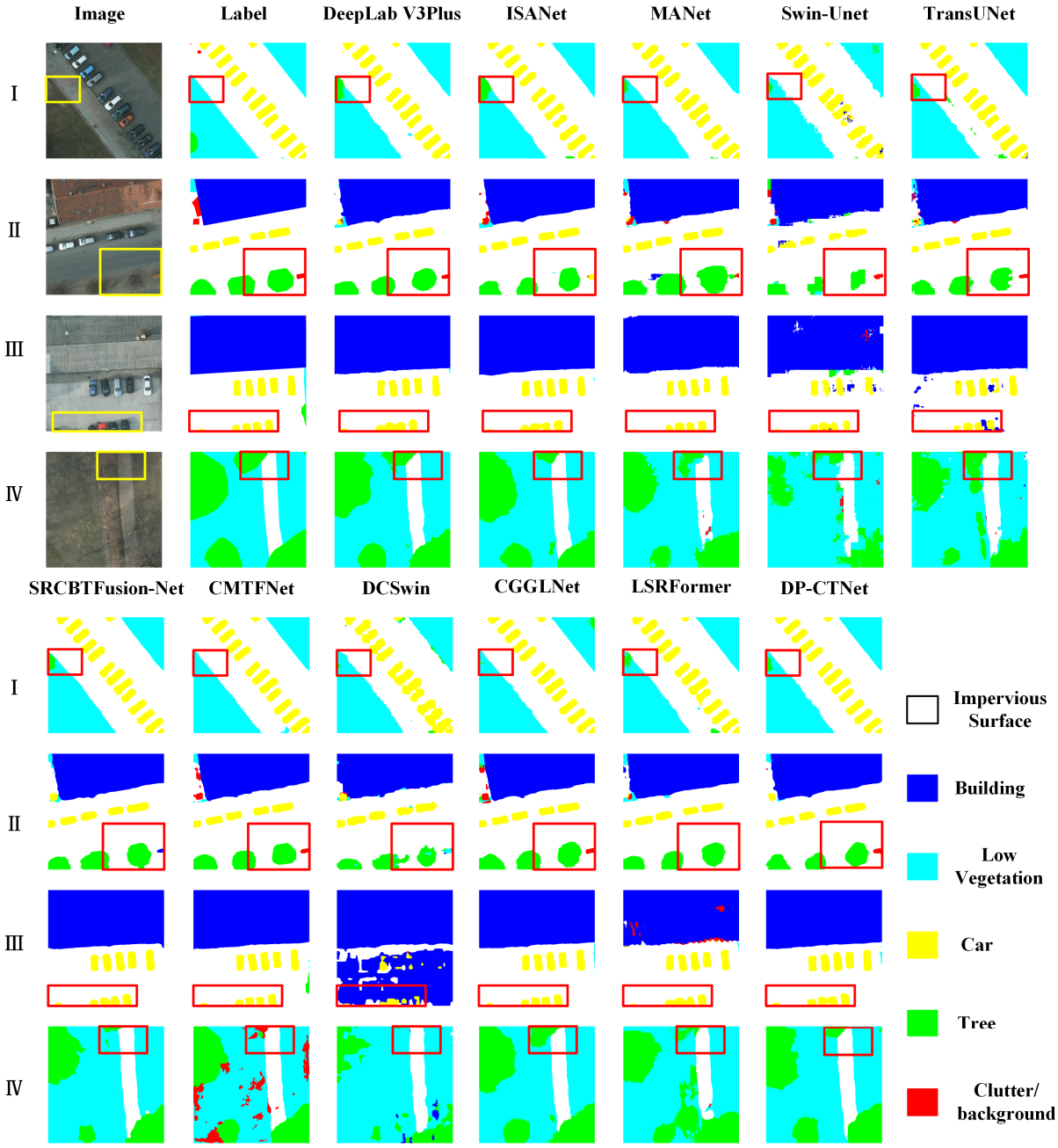


Fig. 4. Examples of Semantic Segmentation Results on the Potsdam Dataset for DP-CTNet and Other Models. The red and yellow boxes highlight noteworthy areas.

TABLE III
COMPARISON FOR DP-CTNET AND OTHER MODELS ON THE POTSDAM DATASET

Models	Parameters Size (MB)	F1-score					OA	MIoU
		Impervious Surface	Building	Low Vegetation	Tree	Car		
DeepLab V3Plus	85.6	88.38%	92.79%	80.80%	80.23%	86.08%	85.78%	75.21%
ISANet	116.83	88.61%	92.93%	81.32%	80.34%	85.29%	86.04%	75.27%
MANet	136.8	87.32%	91.47%	79.55%	78.08%	86.44%	84.45%	73.60%
TransUNet	254.88	84.64%	87.37%	77.40%	76.93%	85.89%	82.08%	70.37%
Swin-Unet	103.67	80.29%	81.94%	71.23%	63.61%	71.73%	75.53%	58.87%
DCSwin	174.07	81.13%	82.62%	72.10%	68.16%	76.15%	78.73%	61.64%
CMTFNet	114.71	88.30%	92.68%	81.35%	81.14%	86.57%	80.97%	75.71%
SRCBTFusion-Net	145.41	88.48%	91.99%	81.00%	81.19%	86.44%	85.05%	75.41%
CGGLNet	62.22	89.36%	89.25%	71.36%	74.41%	87.49%	84.69%	70.77%
LSRFormer	68.03	86.31%	83.30%	65.15%	62.68%	82.70%	80.99%	62.35%
DP-CTNet	196.76	88.59%	**93.09%**	81.02%	**81.40%**	**87.93%**	**86.32%**	**76.36%**

2) Result on Potsdam Dataset: From Table III, it is evident that DP-CTNet exhibits the highest F1-scores for Building, Tree, and Car. Regarding Impervious Surface, DP-CTNet is slightly worse than CGGLNet and ISANet. The performance for Low Vegetation is only slightly lower than that of ISANet, with differences of just 0.3%, respectively. This to some extent suggests that DP-CTNet excels not only in the segmentation accuracy of small-scale objects like Cars but also performs well in the segmentation of large-scale objects like Buildings. In terms of OA and MIoU, the table clearly indicates that DP-CTNet achieves the highest precision, with scores of 86.32% and 76.36%, followed by ISANet and SRCBTFusion-Net. Furthermore, compared to TransUNet and Swin-Unet, CMTFNet, SRCBTFusion-Net, and DP-CTNet demonstrate superior segmentation accuracy, suggesting that these methods are more suitable for semantic segmentation of high-resolution remote sensing images.

Based on the visual analysis of the results in Fig. 4, several observations can be noted. In the first set of results, within the red boxes, there is no presence of Trees. However, different models demonstrate confusion between the Tree and Low Vegetation classes. Notably, DP-CTNet, CMTFNet, CGGLNet, and DCSwin exhibit excellent overall segmentation performance, with minimal confusion between Tree and Low Vegetation, but face issues with fragmented Impervious Surface segmentation in the case of CMTFNet and DCSwin. In the second set of results, only DeepLab V3Plus, DP-CTNet, SRCBTFusion-Net, CGGLNet, LSRFormer, and CMTFNet effectively segment the Tree objects at the bottom, while other models experience the loss of object information. CGGLNet and LSRFormer show some degree of misclassification of Impervious Surface with Clutter/background. Both CMTFNet and DeepLab V3Plus encounter prominent land cover segmentation confusion in the top-left corner. SRCBTFusion-Net misclassifies Buildings within the rectangular box. In the third set, the red box encloses closely spaced Cars affected by shadows, presenting a challenge for segmentation. DP-CTNet, MANet, CGGLNet, and CMTFNet successfully distinguish between individual Cars, while results from other models appear coarse. CMTFNet also demonstrates good segmentation performance for Tree objects on the right edge. LSRFormer has a Building split fragmentation issue. In the fourth set, within the red box, there is a tendency for confusion between Tree, Impervious Surface, and Low Vegetation. SRCBTFusion-Net, CGGLNet and ISANet exhibit overall good segmentation results; however, DP-CTNet outperform other models in this aspect. Overall, DP-CTNet demonstrates exceptional performance in segmenting objects of varying scales.

Based on the comparative results presented above, the DP-CTNet has achieved commendable performance, which can be attributed in part to the model's effective integration of the advantages of both Transformers and CNNs. It optimizes the Transformer's superior capability to capture long-range information while preserving the CNN's ability to finely extract local details from remote sensing imagery.

D. Comparison between EDP-CTNet and other methods

In this paper, we compare EDP-CTNet with existing models, namely BEDSN [77], MGCNet [78], BANet[60], DFANet[50], SegFormer[59], MobileViT-S[30], UNetFormer[33] and Lawin Transformer[61]. These models are all lightweight, with small parameter sizes.

1) Result on Vaihingen Dataset: According to Table IV, EDP-CTNet exhibits a relatively small parameter size, slightly larger than DFANet, at 10.98MB. Concerning F1-score, EDP-CTNet demonstrates outstanding accuracy for various land cover categories, with the exception of Trees, which does not perform as well as BEDSN, SegFormer and BANet. Notably, EDP-CTNet shows a significantly higher F1-score for Cars, surpassing other models by as much as 17.31%. This indicates that EDP-CTNet inherits the excellent capability of DP-CTNet in learning from small-scale objects. In terms of OA and MIoU, EDP-CTNet is still the best-performing model, reaching 85.79% and 70.03% respectively, slightly higher than MSGCNet and much larger than the other models. Thus, the results suggest that EDP-CTNet, after AKD training, maintains impressive accuracy across various land cover categories, while outperforming other lightweight models in overall performance.

Analysing the results from Fig. 5, in the first set of images, EDP-CTNet exhibits some shortcomings in overall Building segmentation within the red box, while SegFormer and MobileViT-S results suffer from overfitting issues. The performance of other models in segmenting Buildings is also less than satisfactory. In the second set of images, within the red box, it's evident that only EDP-CTNet successfully segments Cars in shadowed areas, while other models perform

TABLE IV
COMPARISON FOR EDP-CTNET AND OTHER MODELS ON THE VAIHINGEN DATASET

Models	Parameters Size (MB)	F1-score					OA	MIoU
		Impervious Surface	Building	Low Vegetation	Tree	Car		
SegFormer	14.19	87.35%	87.74%	81.46%	79.42%	58.10%	84.02%	66.25%
Lawin Transformer	17.48	86.29%	86.22%	80.94%	78.15%	57.69%	83.01%	64.86%
DFANet	8.21	87.21%	89.49%	80.93%	79.01%	52.42%	84.19%	65.42%
BANet	16.56	88.29%	86.60%	81.47%	80.28%	59.25%	84.15%	66.65%
MobileViT-S	20.48	87.21%	91.26%	82.53%	78.09%	49.14%	84.92%	65.62%
UNetFormer	44.58	86.56%	88.34%	80.92%	77.86%	57.70%	83.60%	65.53%
MSGCNet	27.02	87.10%	90.42%	83.22%	79.02%	70.51%	85.20%	70.01%
BEDSN	21.27	85.79%	89.14%	84.18%	79.24%	63.46%	84.85%	68.06%
EDP-CTNet	10.98	**88.36%**	**91.58%**	**83.27%**	79.12%	**66.45%**	**85.79%**	**70.03%**

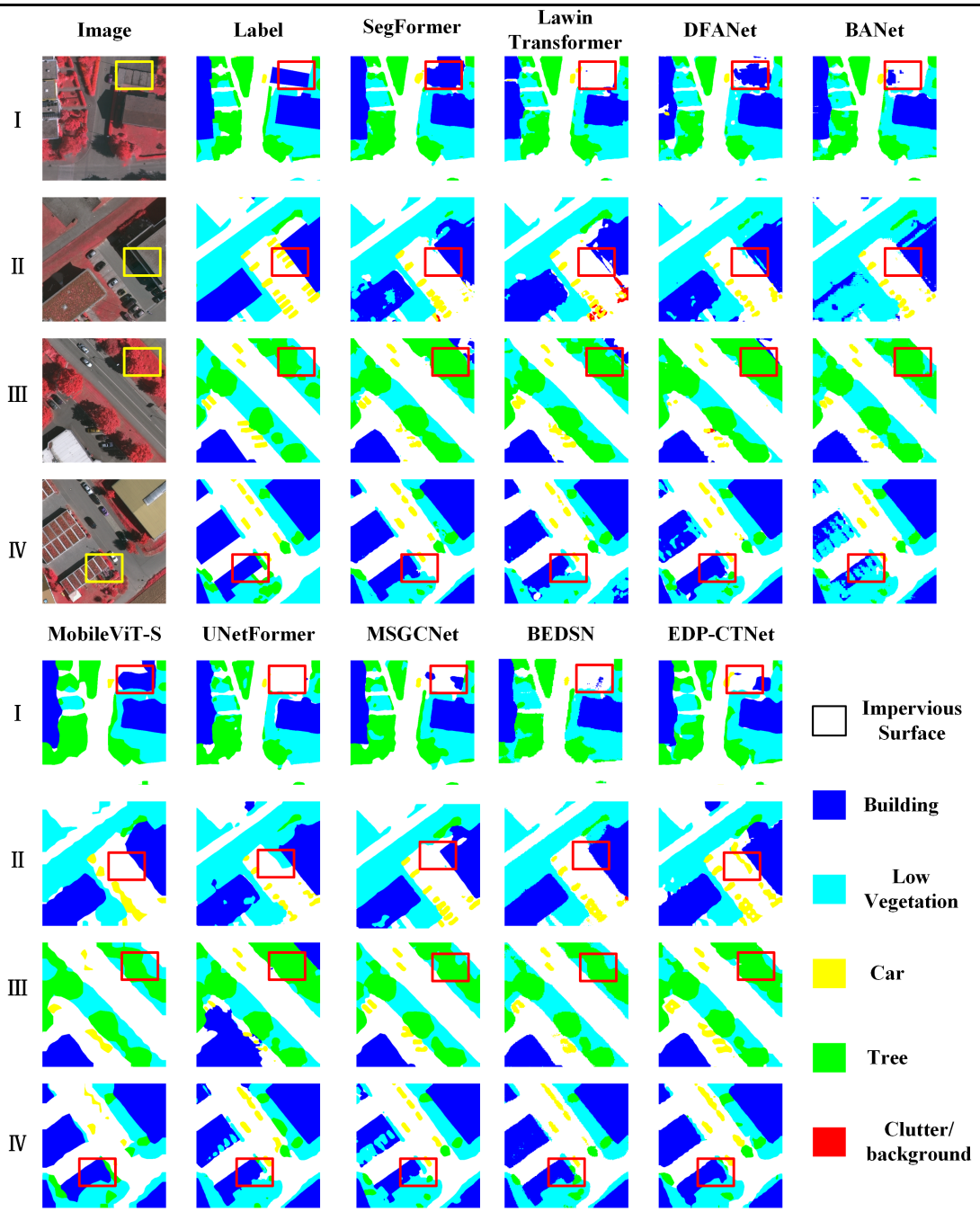


Fig. 5. Examples of Semantic Segmentation Results on the Vaihingen Dataset for EDP-CTNet and Other Models. The red and yellow boxes highlight noteworthy areas.

TABLE V
COMPARISON FOR EDP-CTNET AND OTHER MODELS ON THE POTSDAM DATASET

Models	Parameters Size (MB)	F1-score					OA	MIoU
		Impervious Surface	Building	low Vegetation	Tree	Car		
SegFormer	14.19	86.20%	90.02%	78.01%	75.52%	84.66%	82.96%	71.12%
Lawin Transformer	17.48	85.34%	89.41%	77.52%	74.95%	82.72%	82.24%	69.80%
DFANet	8.21	78.51%	81.71%	67.98%	55.19%	65.80%	72.81%	54.47%
BANet	16.56	83.31%	86.34%	73.99%	67.58%	83.00%	78.73%	65.61%
MobileVIT-S	20.48	87.25%	91.82%	80.85%	80.23%	77.30%	83.20%	70.49%
UNetFormer	44.58	82.91%	84.57%	66.53%	66.34%	79.68%	76.61%	61.96%
MSGCNet	27.02	87.44%	86.80%	68.06%	72.74%	87.83%	83.34%	68.28%
BEDSN	21.27	89.92%	88.55%	71.94%	73.25%	87.50%	82.38%	70.58%
EDP-CTNet	10.98	**88.01%**	**94.41%**	75.71%	**80.78%**	83.43%	**83.38%**	**72.11%**

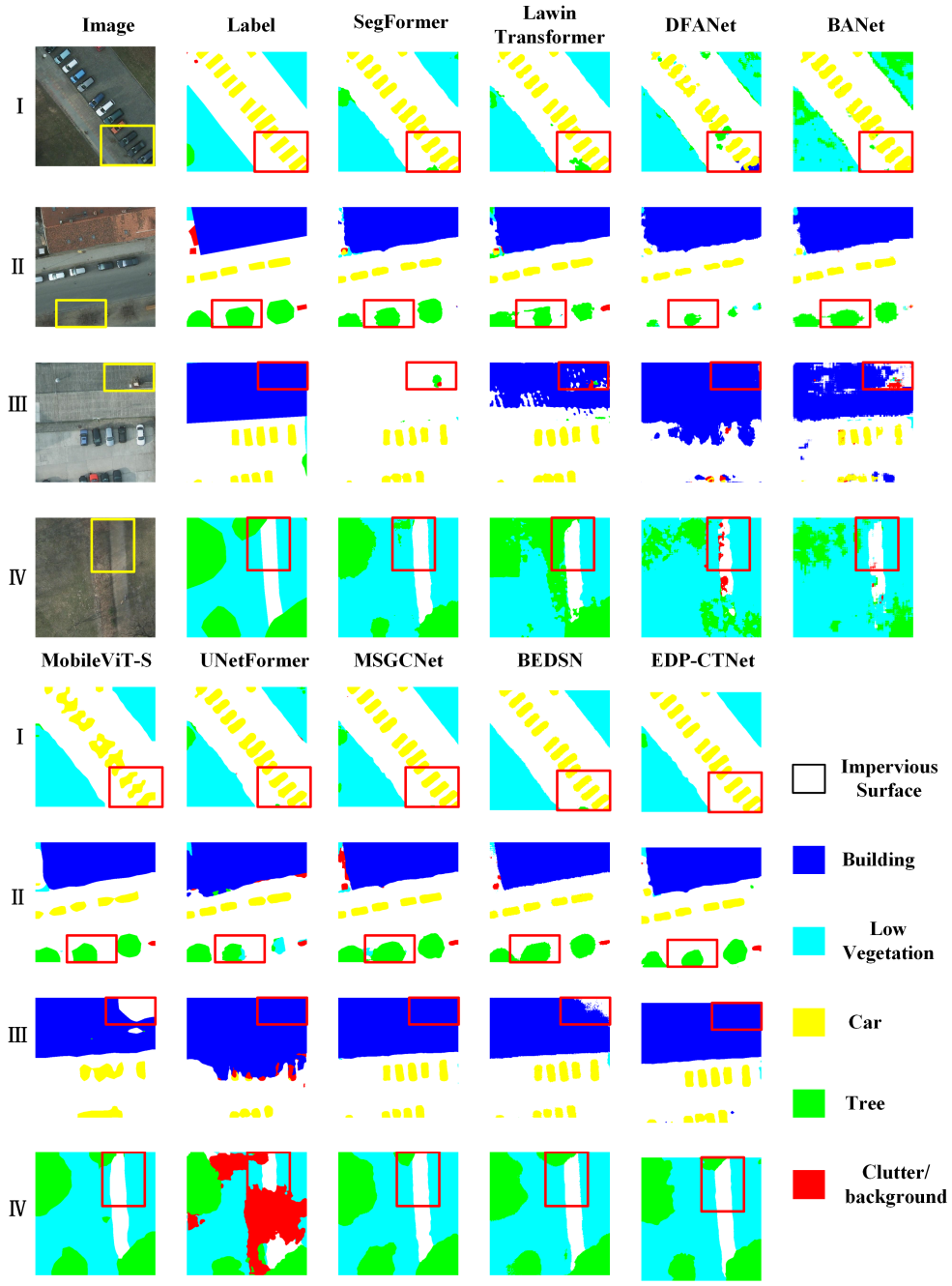


Fig. 6. Examples of Semantic Segmentation Results on the Potsdam Dataset for EDP-CTNet and Other Models. The red and yellow boxes highlight noteworthy areas.

TABLE VI
COMPARISON OF DIFFERENT METHODS IN TERMS OF EFFICIENCY AND ACCURACY

Models	FLOPs (G)	FPS (s)	Potsdam		Vaihingen	
			MIoU	OA	MIoU	OA
DeepLabV3Plus	31.74	121.48	75.21%	85.78%	68.27%	86.48%
ISANet	155.47	21.37	75.27%	86.04%	68.20%	85.97%
MANet	77.83	36.27	73.60%	84.45%	70.69%	86.54%
TransUNet	130.43	23.41	70.37%	82.08%	62.33%	81.97%
Swin-Unet	31.05	42.35	58.87%	75.53%	64.91%	83.97%
DCSwin	46.93	44.32	61.64%	78.73%	62.01%	80.72%
CMTFNet	33.07	42.05	75.71%	80.97%	67.97%	83.24%
SRCBTFusion-Net	77.68	26.55	75.41%	85.05%	70.43%	85.47%
CGGLNet	172.59	26.57	70.77%	84.69%	72.76%	86.03%
LSRFormer	70.91	26.57	62.35%	80.99%	68.99%	85.54%
DP-CTNet	87.51	35.24	76.36%	86.32%	73.03%	86.96%
SegFormer	6.79	89.72	71.12%	82.96%	66.25%	84.02%
Lawin Transformer	7.93	35.05	69.80%	82.24%	64.86%	83.01%
DFANet	1.79	43.47	54.47%	72.81%	65.42%	84.19%
BANet	13.13	34.49	65.61%	78.73%	66.65%	84.15%
MobileViT-S	8.57	87.46	70.49%	83.20%	65.62%	84.92%
UNetFormer	11.74	102.93	61.96%	76.61%	65.53%	83.60%
MSGCNet	28.62	53.29	68.28%	83.34%	70.01%	85.20%
BEDSN	73.46	52.18	70.58%	82.38%	68.06%	84.85%
EDP-CTNet	9.58	105.18	72.11%	83.38%	70.03%	85.79%

poorly, with little or no recognition of Cars. UNetFormer misclassifies these shadowed areas as Low Vegetation. This indicates that EDP-CTNet excels in learning small-scale objects within complex scenes. In the third set of images, EDP-CTNet and MobileViT-S demonstrate superior capabilities in distinguishing Trees and Low Vegetation. MobileViT-S, BEDSN and MSGCNet perform relatively better in segmentation, while other models exhibit significant confusion between Tree and Low Vegetation. In the fourth set of images, within the red box, BANet and Lawin Transformer encounter segmentation fragmentation issues in Building regions, while other models effectively segment Buildings. However, UNetFormer, SegFormer and DFANet misclassify Trees as Low Vegetation, rather than correctly identifying them. Only MobileViT-S and EDP-CTNet maintain the overall continuity of Building segmentation with fewer misclassification issues. Overall, while MobileViT-S demonstrates good segmentation results, EDP-CTNet exhibits superior small-scale object segmentation capabilities. Additionally, EDP-CTNet retains contextual semantic information, ensuring the overall continuity of segmentation results.

2) Result on Potsdam Dataset: From Table V, it is evident that EDP-CTNet exhibits the highest F1-scores for Building, and Tree classes, while delivering moderate performance for Low Vegetation. DP-CTNet's accuracy in segmenting Low Vegetation is similarly modest. Furthermore, EDP-CTNet's F1-score for Car class is lower than MSGCNet and BEDSN. In terms of OA and MIoU, EDP-CTNet outperforms all other models, achieving values of 83.38% and 72.11%, respectively. MSGCNet has slightly lower OA than EDP-CTNet. Overall, EDP-CTNet, trained using AKD, preserves the excellent performance of DP-CTNet in terms of accuracy while surpassing the performance of other models.

Analysing the results from Fig. 6, in the first group, UNetFormer shows a few instances of misclassifying Imperious Surface as Tree. Both MSGCNet, MobileViT-S and EDP-CTNet exhibit no misclassification of Impervious Surface and other land cover categories within the red bounding box. However, EDP-CTNet outperforms MobileViT-S in segmenting Car. In the second group, within the red bounding box, SegFormer, EDP-CTNet, and MobileViT-S demonstrate commendable segmentation of Tree, while MobileViT-S still faces challenges in Car segmentation, and other models exhibit fragmentation issues. BEDSN and MSGCNet appear to be connected between the Tree. In the third group, within the red bounding box, the segmentation results of UNetFormer are comparatively poor. EDP-CTNet displays the most cohesive overall performance in Building segmentation, while other models manifest discontinuities, with SegFormer showing the weakest performance. In the fourth group, relative to the other models, EDP-CTNet effectively distinguishes between Tree, Impervious Surface, and Low Vegetation. In summary, with the assistance of AKD, EDP-CTNet maintains excellent segmentation performance while improving inference speed.

The comparative results demonstrate that the EDP-CTNet not only achieves a lightweight architecture but also maintains superior segmentation efficacy. This is attributable to two main factors: first, the employment of structured pruning and the transition to depthwise separable convolutions, which significantly reduce the number of model parameters; second, the integration of AKD, which ensures that the EDP-CTNet inherits the segmentation capabilities of the teacher network effectively.

E. Efficiency Analysis

To demonstrate that EDP-CTNet can improve the inference

speed of remote sensing image semantic segmentation to some extent, this study utilizes the frames per second (FPS) metric for comparing the inference speed of different models under identical experimental conditions. A higher FPS value indicates a faster model inference speed. As shown in Fig. 12, the fastest inference speed is achieved by DeepLab V3Plus due to its relative simplicity compared to transformer models. Given the relatively complex Swin Transformer block in DP-CTNet, its inference speed is moderate, surpassing only ISANet and TransUNet. The underlying cause for this stems from the high complexity and memory access costs incurred by the self-attention mechanism of Transformers, as well as the substantial number of parameters introduced by the multiple upsampling and fusion stages within the architecture. However, as indicated in Table II, the DP-CTNet surpasses the performance of other models.

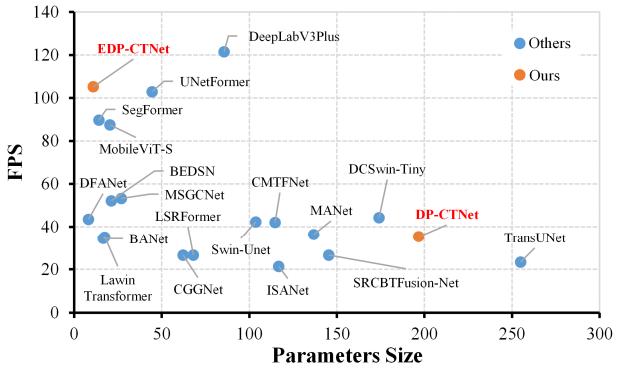


Fig. 12. Comparison of FPS Among Different Models.

Furthermore, EDP-CTNet's inference speed significantly improves, ranking just below DeepLab V3Plus, ranking just below DeepLab V3Plus, and it is faster than other transformer-based models, with a parameter size of only 10.98MB. This suggests that the employment of depthwise separable convolutions and structured pruning significantly reduces the number of parameters, thereby enhancing the model's inference speed.

In order to prove that the segmentation accuracy of EDP-CTNet is not bad while keeping lightweight, this paper calculates the accuracy metrics and FPS and FLOPs of different models, as in Table VI. FPS can reflect the inference speed of the model, and FLOPs can reflect the computation amount of the model. From the table, it can be seen that DeepLabV3 Plus, DP-CTNet and SRCBTFusion-Net maintain a good balance of computation, inference speed and accuracy, especially in the Potsdam dataset. EDP-CTNet has relatively low FLOPs and a high FPS of 105.18, which suggests that it is able to achieve fast inference speeds while maintaining a low computation requirement while being able to achieve fast inference speeds. Although the accuracy is not as good as some more complex models, it is still a competitive choice compared to other lightweight models.

VI. Conclusion

In the realm of remote sensing image semantic segmentation, the Transformer architecture has gained prominence as a leading approach. However, it still grapples with issues like slower inference speeds and potential loss of fine-grained details. We propose DP-CTNet, a model that features a dual pathway, combining both CNN and Transformer components. Additionally, we design two novel modules: FRM for guiding Transformer feature learning and FFM to effectively merge CNN and Transformer features. DP-CTNet undergoes a pruning process to create a lightweight variant known as EDP-CTNet. To enhance its segmentation performance and speed, we employ the knowledge distillation technique known as AKD during the training of EDP-CTNet. Experimental results affirm that DP-CTNet excels at capturing long contextual semantic information while preserving fine-grained details of small-scale objects. EDP-CTNet, trained with AKD, significantly improves segmentation speed without compromising accuracy compared to other models. However, it's important to note that both DP-CTNet and EDP-CTNet exhibit extended training times and weaker performance in edge segmentation. In the upcoming phases of our research, we aim to overcome these limitations and delve into the development of a remote sensing image semantic segmentation model that minimizes overall time costs and exhibits robust generalization capabilities within the constraints of limited training samples.

References

[1] X. Li *et al.*, 'Big Data in Earth system science and progress towards a digital twin', *Nat Rev Earth Environ*, vol. 4, no. 5, pp. 319–332, May 2023, doi: 10.1038/s43017-023-00409-w.

[2] J. Gong, C. Liu, and X. Huang, 'Advances in urban information extraction from high-resolution remote sensing imagery', *Sci. China Earth Sci.*, vol. 63, no. 4, pp. 463–475, Apr. 2020, doi: 10.1007/s11430-019-9547-x.

[3] P. Gong *et al.*, 'Mapping essential urban land use categories in China (EULUC-China): preliminary results for 2018', *Science Bulletin*, vol. 65, no. 3, pp. 182–187, Feb. 2020, doi: 10.1016/j.scib.2019.12.007.

[4] X. Huang and Y. Liu, 'Livability assessment of 101,630 communities in China's major cities: A remote sensing perspective', *Science China Earth Sciences*, vol. 65, no. 6, pp. 1073–1087, Jun. 2022, doi: 10.1007/s11430-021-9896-4.

[5] S. Peng *et al.*, 'Surface Urban Heat Island Across 419 Global Big Cities', *Environ. Sci. Technol.*, vol. 46, no. 2, pp. 696–703, Jan. 2012, doi: 10.1021/es2030438.

[6] M. Reichstein *et al.*, 'Deep learning and process understanding for data-driven Earth system science', *Nature*, vol. 566, no. 7743, pp. 195–204, Feb. 2019, doi: 10.1038/s41586-019-0912-1.

[7] X. Yuan, J. Shi, and L. Gu, 'A review of deep learning methods for semantic segmentation of remote sensing imagery', *Expert Systems with Applications*, vol. 169, p. 114417, May 2021, doi: 10.1016/j.eswa.2020.114417.

[8] E. Adam, O. Mutanga, J. Odindi, and E. M. Abdel-Rahman, 'Land-use/cover classification in a heterogeneous coastal landscape using RapidEye

imagery: evaluating the performance of random forest and support vector machines classifiers', *International Journal of Remote Sensing*, vol. 35, no. 10, pp. 3440–3458, May 2014, doi: 10.1080/01431161.2014.903435.

[9] T. Ishida *et al.*, 'A novel approach for vegetation classification using UAV-based hyperspectral imaging', *Computers and Electronics in Agriculture*, vol. 144, pp. 80–85, Jan. 2018, doi: 10.1016/j.compag.2017.11.027.

[10] B. Chen *et al.*, 'Mapping essential urban land use categories with open big data: Results for five metropolitan areas in the United States of America', *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 203–218, Aug. 2021, doi: 10.1016/j.isprsjprs.2021.06.010.

[11] K. Zheng, H. Zhang, H. Wang, F. Qin, Z. Wang, and J. Zhao, 'Using Multiple Sources of Data and "Voting Mechanisms" for Urban Land-Use Mapping', *Land*, vol. 11, no. 12, Art. no. 12, Dec. 2022, doi: 10.3390/land11122209.

[12] K. Zheng, H. Wang, F. Qin, and Z. Han, 'A Land Use Classification Model Based on Conditional Random Fields and Attention Mechanism Convolutional Networks', *Remote Sensing*, vol. 14, no. 11, Art. no. 11, Jan. 2022, doi: 10.3390/rs14112688.

[13] K. Zheng, H. Wang, F. Qin, C. Miao, and Z. Han, 'An improved land use classification method based on DeepLab V3+ under GauGAN data enhancement', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 5526–5537, 2023, doi: 10.1109/JSTARS.2023.3278862.

[14] O. Ronneberger, P. Fischer, and T. Brox, 'U-Net: Convolutional Networks for Biomedical Image Segmentation', in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, 'Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation', in *Computer Vision – ECCV 2018*, vol. 11211, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science, vol. 11211. , Cham: Springer International Publishing, 2018, pp. 833–851. doi: 10.1007/978-3-030-01234-2_49.

[16] H. Jung, H.-S. Choi, and M. Kang, 'Boundary Enhancement Semantic Segmentation for Building Extraction From Remote Sensed Image', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022, doi: 10.1109/TGRS.2021.3108781.

[17] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[18] X. Yang *et al.*, 'An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery', *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 177, pp. 238–262, Jul. 2021, doi: 10.1016/j.isprsjprs.2021.05.004.

[19] X. Chen *et al.*, 'Adaptive Effective Receptive Field Convolution for Semantic Segmentation of VHR Remote Sensing Images', *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 4, pp. 3532–3546, Apr. 2021, doi: 10.1109/TGRS.2020.3009143.

[20] I. Kotaridis and M. Lazaridou, 'Remote sensing image segmentation advances: A meta-analysis', *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 309–322, Mar. 2021, doi: 10.1016/j.isprsjprs.2021.01.020.

[21] A. Vaswani *et al.*, 'Attention is all you need', in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 6000–6010.

[22] A. Dosovitskiy *et al.*, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', in *International Conference on Learning Representations*, 2022.

[23] Z. Liu *et al.*, 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[24] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, 'A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images', *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3143368.

[25] L. Gao *et al.*, 'STransFuse: Fusing Swin Transformer and Convolutional Neural Network for Remote Sensing Image Semantic Segmentation', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10990–11003, 2021, doi: 10.1109/JSTARS.2021.3119654.

[26] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, 'Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022, doi: 10.1109/TGRS.2022.3144165.

[27] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, 'Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022, doi: 10.1109/TGRS.2022.3144894.

[28] Y. Liu, Y. Zhang, Y. Wang, and S. Mei, 'Rethinking Transformers for Semantic Segmentation of Remote Sensing Images', *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–15, 2023, doi: 10.1109/TGRS.2023.3302024.

[29] J. Chen, J. Yi, A. Chen, and H. Lin, 'SRCBTFusion-Net: An Efficient Fusion Architecture via Stacked Residual Convolution Blocks and Transformer for Remote Sensing Image Semantic Segmentation', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023, doi: 10.1109/TGRS.2023.3336689.

[30] S. Mehta and M. Rastegari, 'MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision

Transformer', presented at the International Conference on Learning Representations, Oct. 2021. Accessed: Sep. 29, 2023. [Online]. Available: https://openreview.net/forum?id=vh-0sUt8HlG

[31] W. Zhang *et al.*, 'TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12083–12093. Accessed: Feb. 29, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Zhang_TopFormer_Token_Pyramid_Transformer_for_Mobile_Semantic_Segmentation_CVPR_2022_paper.html?ref=https://githubhelp.com

[32] J. Zheng, L. Yang, Y. Li, K. Yang, Z. Wang, and J. Zhou, 'Lightweight Vision Transformer with Spatial and Channel Enhanced Self-Attention', presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 1492–1496. Accessed: Feb. 29, 2024. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023W/RCV/html/Zheng_Lightweight_Vision_Transformer_with_Spatial_and_Channel_Enhanced_Self-Attention_ICCVW_2023_paper.html

[33] L. Wang *et al.*, 'UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery', *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, Aug. 2022, doi: 10.1016/j.isprsjprs.2022.06.008.

[34] Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, 'Efficient Transformer for Remote Sensing Image Segmentation', *Remote Sensing*, vol. 13, no. 18, Art. no. 18, Jan. 2021, doi: 10.3390/rs13183585.

[35] Y. Chen *et al.*, 'LightFGCNet: A Lightweight and Focusing on Global Context Information Semantic Segmentation Network for Remote Sensing Imagery', *Remote Sensing*, vol. 14, no. 24, Art. no. 24, Jan. 2022, doi: 10.3390/rs14246193.

[36] R. Guan, M. Wang, L. Bruzzone, H. Zhao, and C. Yang, 'Lightweight Attention Network for Very High-Resolution Image Semantic Segmentation', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023, doi: 10.1109/TGRS.2023.3272614.

[37] X. Lin, L. Yu, K.-T. Cheng, and Z. Yan, 'The Lighter the Better: Rethinking Transformers in Medical Image Segmentation Through Adaptive Pruning', *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2325–2337, Aug. 2023, doi: 10.1109/TMI.2023.3247814.

[38] W. Zhou, X. Fan, W. Yan, S. Shan, Q. Jiang, and J.-N. Hwang, 'Graph Attention Guidance Network With Knowledge Distillation for Semantic Segmentation of Remote Sensing Images', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023, doi: 10.1109/TGRS.2023.3311480.

[39] Z. Li *et al.*, 'When Object Detection Meets Knowledge Distillation: A Survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10555–10579, Aug. 2023, doi: 10.1109/TPAMI.2023.3257546.

[40] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, 'BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation', in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 334–349. doi: 10.1007/978-3-030-01261-8_20.

[41] J. Gou, B. Yu, S. J. Maybank, and D. Tao, 'Knowledge Distillation: A Survey', *Int J Comput Vis*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021, doi: 10.1007/s11263-021-01453-z.

[42] E. Shelhamer, J. Long, and T. Darrell, 'Fully Convolutional Networks for Semantic Segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, Art. no. 4, Apr. 2017, doi: 10.1109/TPAMI.2016.2572683.

[43] J. Long, E. Shelhamer, and T. Darrell, 'Fully Convolutional Networks for Semantic Segmentation', presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440. Accessed: Oct. 19, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

[44] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778. Accessed: May 14, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, 'DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, Art. no. 4, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[46] B. Huang, B. Zhao, and Y. Song, 'Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery', *Remote Sensing of Environment*, vol. 214, pp. 73–86, Sep. 2018, doi: 10.1016/j.rse.2018.04.050.

[47] Y. Sun, X. Zhang, Q. Xin, and J. Huang, 'Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data', *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 143, pp. 3–14, Sep. 2018, doi: 10.1016/j.isprsjprs.2018.06.005.

[48] R. Li *et al.*, 'Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022, doi: 10.1109/TGRS.2021.3093977.

[49] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang, 'Interlaced Sparse Self-Attention for Semantic

Segmentation', Jul. 30, 2019, *arXiv*: arXiv:1907.12273. doi: 10.48550/arXiv.1907.12273.

[50] H. Li, P. Xiong, H. Fan, and J. Sun, 'DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation', in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 9514–9523. doi: 10.1109/CVPR.2019.00975.

[51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *2019 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES (NAACL HLT 2019), VOL. 1*. ASSOC COMPUTATIONAL LINGUISTICS-ACL, 209 N EIGHTH STREET, STROUDSBURG, PA 18360 USA, pp. 4171–4186, 2019.

[52] K. Han *et al.*, 'A Survey on Vision Transformer', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[53] Y. Zhou, S. Chen, J. Zhao, R. Yao, Y. Xue, and A. E. Saddik, 'CLT-Det: Correlation Learning Based on Transformer for Detecting Dense Objects in Remote Sensing Images', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022, doi: 10.1109/TGRS.2022.3204770.

[54] L. Peng, C. Zhu, and L. Bian, 'U-Shape Transformer for Underwater Image Enhancement', *IEEE Transactions on Image Processing*, vol. 32, pp. 3066–3079, 2023, doi: 10.1109/TIP.2023.3276332.

[55] Y. Liu *et al.*, 'A Survey of Visual Transformers', *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2023, doi: 10.1109/TNNLS.2022.3227717.

[56] H. Cao *et al.*, 'Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation', in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds., in Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 205–218. doi: 10.1007/978-3-031-25066-8_9.

[57] J. Chen *et al.*, 'TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation', Feb. 08, 2021, *arXiv*: arXiv:2102.04306. Accessed: Jul. 22, 2022. [Online]. Available: http://arxiv.org/abs/2102.04306

[58] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, 'CMTFNet: CNN and Multiscale Transformer Fusion Network for Remote-Sensing Image Semantic Segmentation', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023, doi: 10.1109/TGRS.2023.3314641.

[59] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, 'SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 12077–12090. Accessed: Sep. 29, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html

[60] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, 'Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images', *Remote Sensing*, vol. 13, no. 16, Art. no. 16, Jan. 2021, doi: 10.3390/rs13163065.

[61] H. Yan, C. Zhang, and M. Wu, 'Lawin Transformer: Improving Semantic Segmentation Transformer with Multi-Scale Representations via Large Window Attention', arXiv.org. Accessed: Sep. 29, 2023. [Online]. Available: https://arxiv.org/abs/2201.01615v4

[62] B. Cheng, A. Schwing, and A. Kirillov, 'Per-Pixel Classification is Not All You Need for Semantic Segmentation', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 17864–17875. Accessed: Sep. 29, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/hash/950a4152c2b4aa3ad78bdd6b366cc179-Abstract.html

[63] W. Wang *et al.*, 'Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 548–558. doi: 10.1109/ICCV48922.2021.00061.

[64] M. D. Zeiler and R. Fergus, 'Visualizing and Understanding Convolutional Networks', in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 818–833. doi: 10.1007/978-3-319-10590-1_53.

[65] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, 'CBAM: Convolutional Block Attention Module', in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 3–19. doi: 10.1007/978-3-030-01234-2_1.

[66] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, 'BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation', *International Journal of Computer Vision*, vol. 129, no. 11, Art. no. 11, Nov. 2021, doi: 10.1007/s11263-021-01515-2.

[67] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, 'Strip Pooling: Rethinking Spatial Pooling for Scene Parsing', in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 4002–4011. doi: 10.1109/CVPR42600.2020.00406.

[68] J. Guo *et al.*, 'CMT: Convolutional Neural Networks Meet Vision Transformers', in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 12165–12175. doi: 10.1109/CVPR52688.2022.01186.

[69] X. Li, W. Wang, X. Hu, and J. Yang, 'Selective Kernel Networks', in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 510–519. doi: 10.1109/CVPR.2019.00060.

[70] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, 'Training data-efficient image transformers & distillation through attention', in

Proceedings of the 38th International Conference on Machine Learning, PMLR, Jul. 2021, pp. 10347–10357. Accessed: Oct. 09, 2023. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html

[71] M. Ponti, J. Kittler, M. Riva, T. de Campos, and C. Zor, 'A decision cognizant Kullback–Leibler divergence', *Pattern Recognition*, vol. 61, pp. 470–478, Jan. 2017, doi: 10.1016/j.patcog.2016.08.018.

[72] E. J. Crowley, G. Gray, and A. Storkey, 'Moonshine: distilling with cheap convolutions', in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, in NIPS'18. Red Hook, NY, USA: Curran Associates Inc., Dec. 2018, pp. 2893–2903.

[73] J. Ba and R. Caruana, 'Do Deep Nets Really Need to be Deep?', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2014. Accessed: Oct. 09, 2023. [Online]. Available: https://papers.nips.cc/paper_files/paper/2014/hash/ea8fcd92d59581717e06eb187f10666d-Abstract.html

[74] G. Hinton, O. Vinyals, and J. Dean, 'Distilling the Knowledge in a Neural Network', Mar. 09, 2015, *arXiv*: arXiv:1503.02531. doi: 10.48550/arXiv.1503.02531.

[75] Y. Ni, J. Liu, W. Chi, X. Wang, and D. Li, 'CGGLNet: Semantic Segmentation Network for Remote Sensing Images Based on Category-Guided Global–Local Feature Interaction', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024, doi: 10.1109/TGRS.2024.3379398.

[76] R. Zhang, Q. Zhang, and G. Zhang, 'LSRFormer: Efficient Transformer Supply Convolutional Neural Networks With Global Information for Aerial Image Segmentation', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024, doi: 10.1109/TGRS.2024.3366709.

[77] X. Li, L. Xie, C. Wang, J. Miao, H. Shen, and L. Zhang, 'Boundary-enhanced dual-stream network for semantic segmentation of high-resolution remote sensing images', *GIScience & Remote Sensing*, vol. 61, no. 1, p. 2356355, Dec. 2024, doi: 10.1080/15481603.2024.2356355.

[78] Q. Zeng, J. Zhou, J. Tao, L. Chen, X. Niu, and Y. Zhang, 'Multiscale Global Context Network for Semantic Segmentation of High-Resolution Remote Sensing Images', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024, doi: 10.1109/TGRS.2024.3393489.

Kang Zheng received his Master of Science degree in Geography and Environment from Henan University in 2023. He is currently pursuing his PhD degree in School of Resources and Environment, Wuhan University.

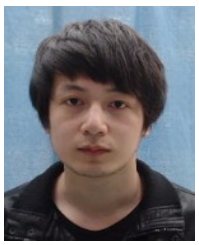
His research interests include GeoAI and GIS.

Yu Chen received the master's degree from Information Engineering University, Zhengzhou, China, in 2017, and the Ph.D. degree in geodesy and surveying engineering from the School of Geodesy and Geomatics, Wuhan University, Wuhan, China, in 2021. He is currently a Lecturer with the School of Geomatics Science and Technology, Nanjing Tech University, Nanjing, China. His research interests include point cloud model reconstruction, multisensory integrated navigation, UAV autonomous route planning, and VIO/VSLAM.

Jingrong Wang received the M.E. degree in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) from Wuhan University, Wuhan, China, in 2018 and 2020, respectively, where, he is currently pursuing the Ph.D degree with the GNSS Research Center. His research interests include GNSS precise positioning, visual inertial odometry (VIO) and multi-sensor fusion algorithm.

Zhifei Liu received the B.E. degree in geographic information science from China University of Petroleum (east China), Qingdao, China, in 2023. He is currently pursuing the M.S. degree with Technical University of Munich, Munich, Germany.

Shuai Bao received the M.S. degree in Cartography and Geographical Information Engineering from Liaoning Technical University, Fuxin, China, in 2023. He is currently pursuing the Ph.D. degree in Resources and Environment at Wuhan University, Wuhan, China.



Jiao Zhan received the M.E. degree in Geomatics Engineering from Wuhan University, Wuhan, China, in 2018. He is currently working toward the Ph.D. degree in Geodesy and Survey Engineering with the Research Center of GNSS, Wuhan University, Wuhan, China. His research interests include Compute Vision, High precision Map and Autonomous Driving.

Nan Shen received the Ph.D. degree from Wuhan University in 2021. He is currently a Lecturer with Nanjing Tech University. His research interests focus on precise GNSS data processing and its applications.